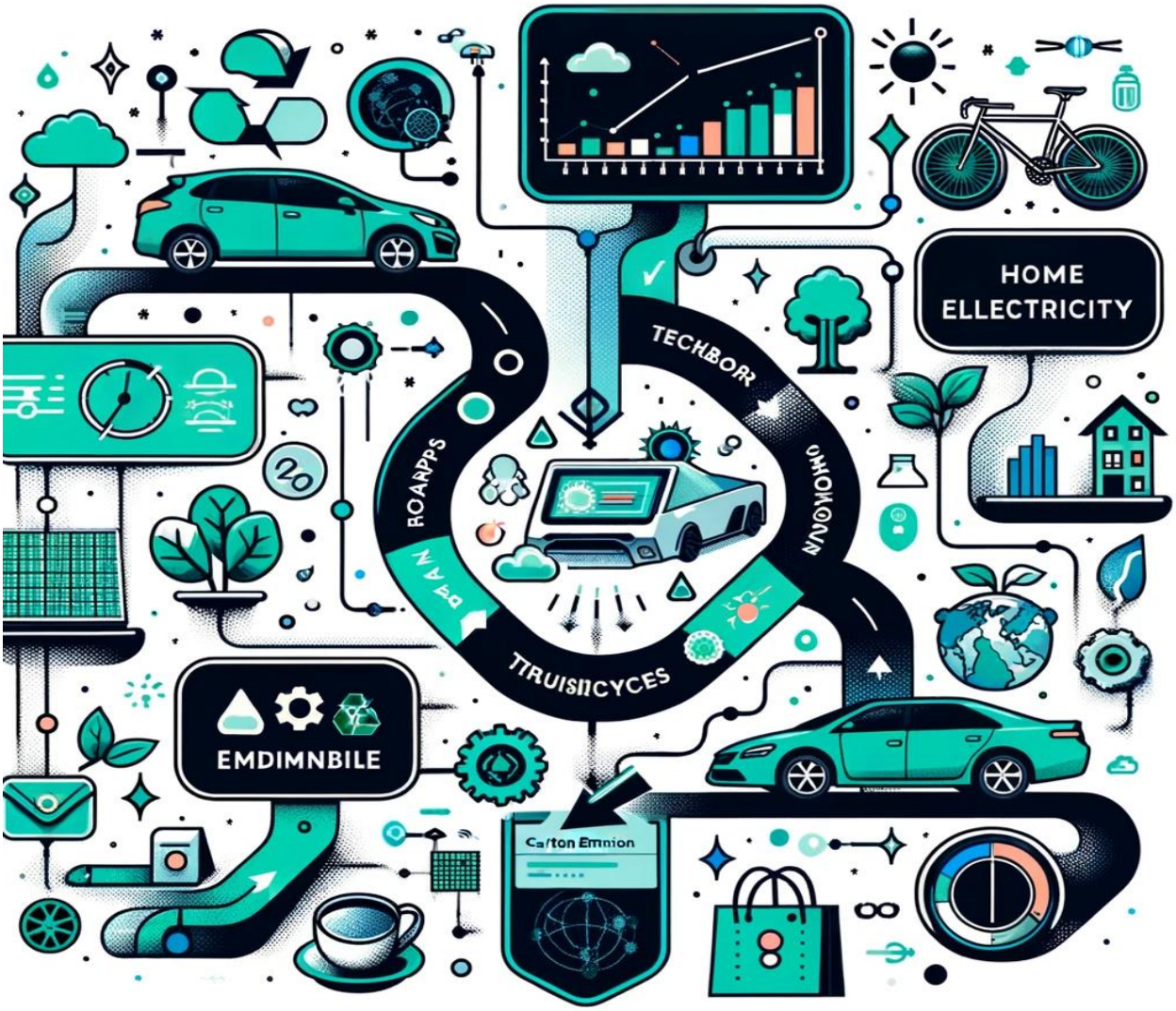
INTRODUCTION TO MACHINE LEARNING

ESTIMATING CARBON EMISSIONS USING MACHINE LEARNING: AN ANALYSIS OF LIFESTYLE FACTOR



PROJECT BY:
ANGAD SINGH

INSTRUCTOR:
ELHAM K. BUXTON

ABSTRACT

In this project we will be investigating the predictive relationship between lifestyle factor and carbon emission using various machine learning models. A diverse dataset which contains transportation, diet, energy usage and other lifestyle choices will be used to train models including the k-nearest neighbors (KNN), Linear regression, regularized regression such as Lasso, Ridge, Elastic Net etc. Support Vector machine (SVM), Random Forest, Gradient boosting, and then finally Neural network. We will use Cross validation and evaluation metrics such as Root Mean Square Error (RMSE), Mean Absolute error (MAE) and Rsquared to evaluate the model's performance. We will be highlighting the ability to capture nonlinear relationship between lifestyle factor and emission. How features like transport energy source can significantly contribute to the carbon emission, we will see how data driven models can potentially emphasize suitable policy and behavioral change.

PROBLEM/BENCHMARK

The main aim of this project is addressing the gap between how an individuals transport choices dietary habits, and energy consumption can significantly impact carbon footprint, by using machine learning models we will be able to see which variables are affecting the most in the carbon emission and we will see that which model will have best outcome.

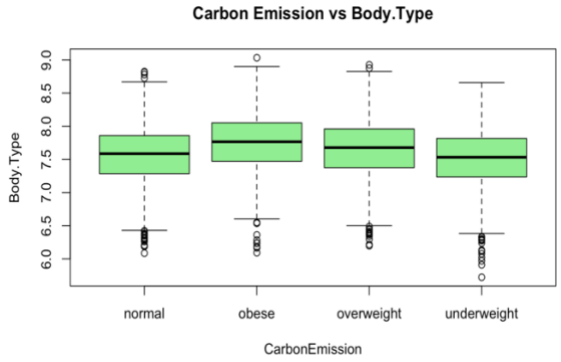
DATASET DESCRIPTION

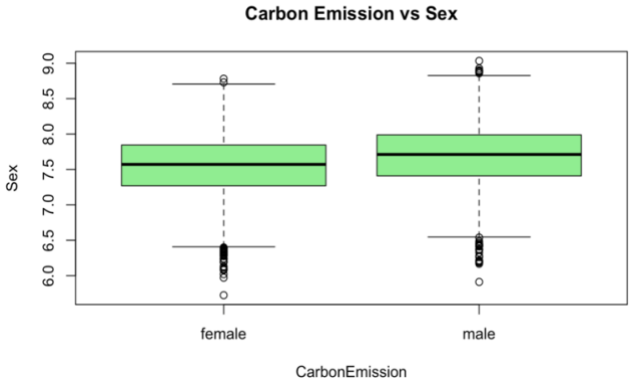
The dataset that I am using in this project is originally named as individual carbon footprint calculation, I got the dataset from Kaggle the dataset had 20 Column's.

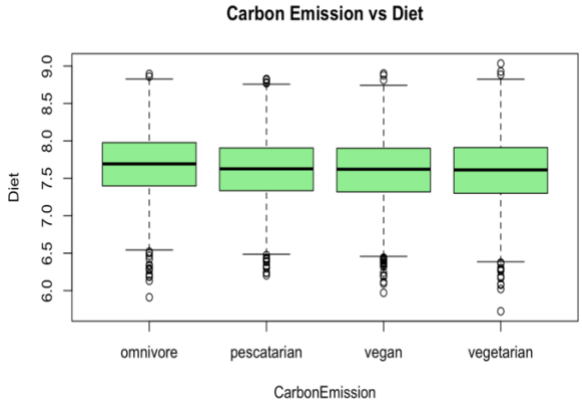
- **Body Type** – it's a categorical variable recording if the person weight comes under normal, obese, overweight, underweight.
- **Sex** - It's a categorical variable having individuals sex Male, Female.
- **Diet** – Categorical variable having if the individual diet is omnivore, pescatarian, vegan, vegetarian**.**
- **How often Shower** – Categorical variable having how often does the individual showers i.e. Daily, less frequently, more Frequently, twice a day.
- **Heating Energy Source** – Categorical variable having how does the individual heats their residential are Coal, Electricity, natural gas, wood.
- **Transport** – Categorical variable storing the transport preference of an individual i.e. private, public, walk/bicycle.
- **Vehicle Type** – Categorical variable which has individual vehicle type diesel, electric, hybrid, lpg, petrol and this column was having NA values which was related with the previous column for the people who used public and walk/bicycle as their transport, they were left with NA in this column, so I made a separate category for them as FuelEfficient**.**
- **Social Activity** – Categorical Column having how often the individual participates in social activity. (never, often, sometimes)
- **Monthly Grocery Bill** – Individuals Monthly grocery bill ranging from 50 to 299.
- **Frequency of Travelling by Air** – Categorical Variable having how often did the individual travel by air in the last month. (frequently, never, rarely, very frequently)
- **Vehicle Monthly Distance Traveled** – How much did the individual travel by vehicle in the previous month.
- **Waste Bag Size** – Categorical variable having what size of garbage bag does the individual use. (small, medium, large, extra-large)
- **Waste Bag Weekly Count** – The amount of garbage bags thrown in last week.
- **How long TV/PC Daily hour** – Daily time Spent Infront of TV or PC.
- **How many new cloths monthly** – Number of cloths did the individual buy in a month.
- **Howhw long internet Daily Hour** – The time spent by an individual on the internet in a day.
- **Energy Efficiency** – Categorical column having if the individual cares about purchasing an energy efficient device. (**No**, **Sometime**, **Yes**)
- **Recycling** – The type of waste an individual recycles. (Metal, Paper, Plastic, Glass)
- **Cooking with** – Device used by an individual for cooking. (Stove, Oven, Microwave, Grill, Air fryer)

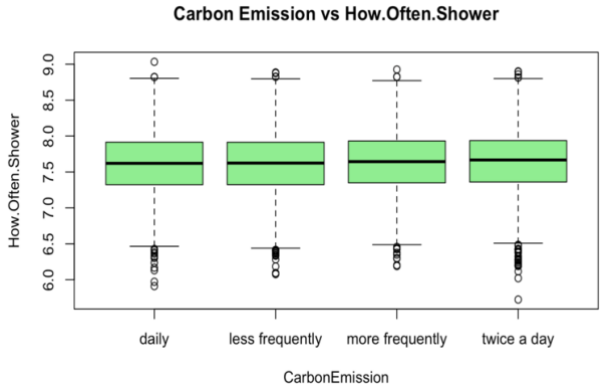
- **Carbon Emission** – Target Variable, Total Carbon Emission caused.

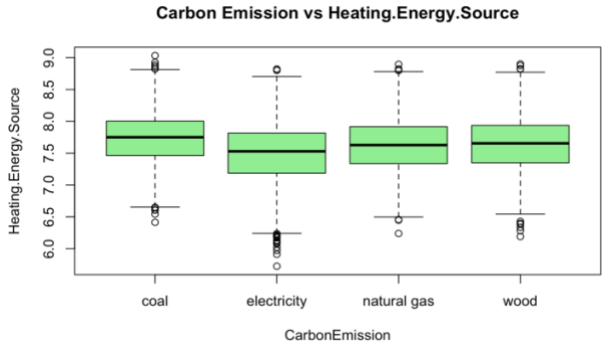
EXPLORATION OF DATA

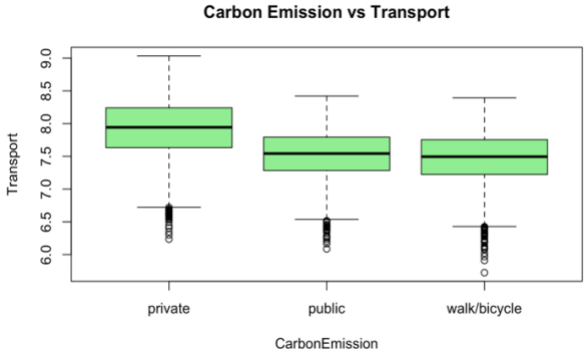


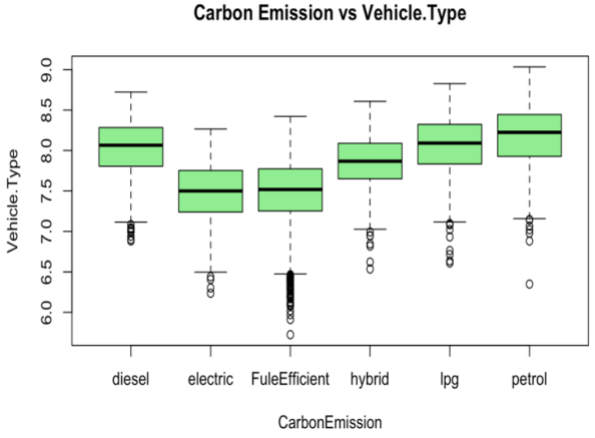


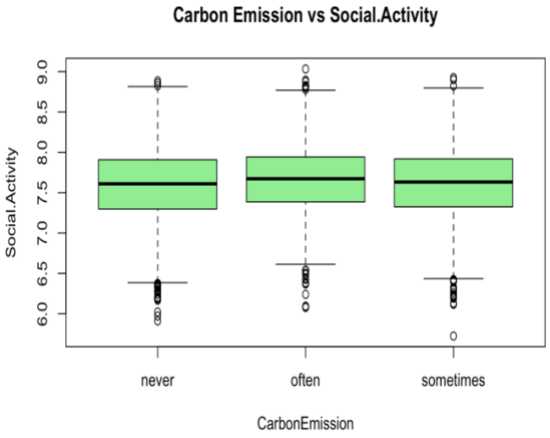


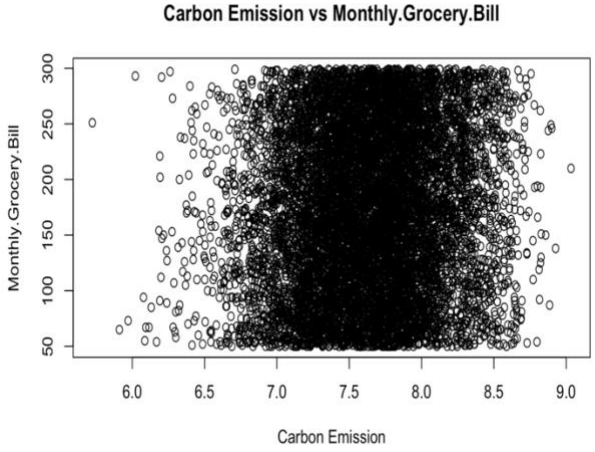


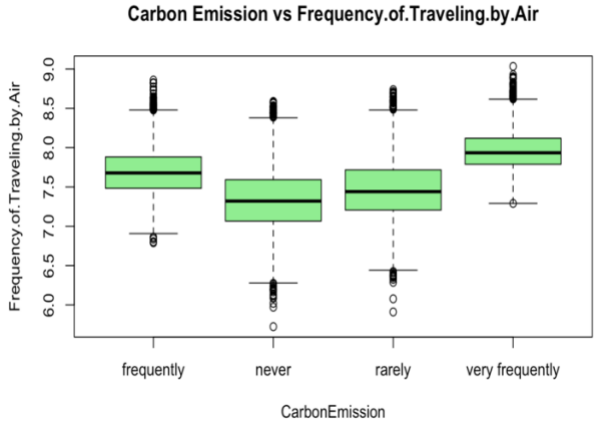


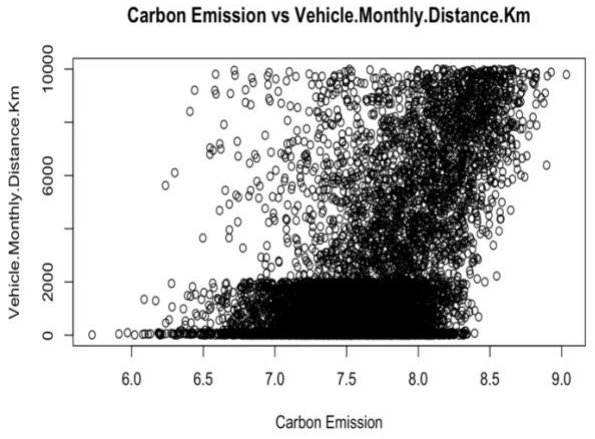


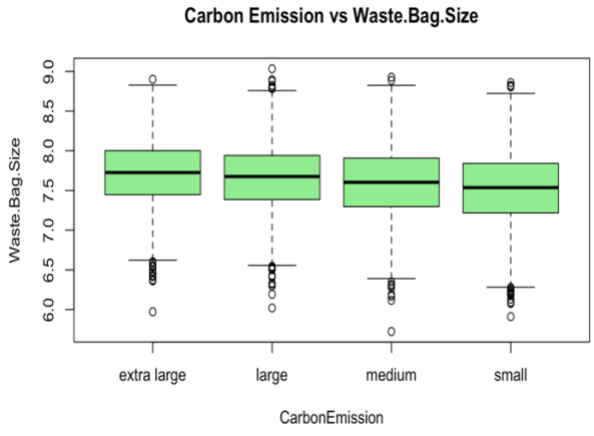

Carbon Emission vs Vehicle.Type
Vehicle.Type
6.0 6.5 7.0 7.5 8.0 8.5 9.0
diesel
electric
FuleEfficient
hybrid
lpg
petrol
CarbonEmission

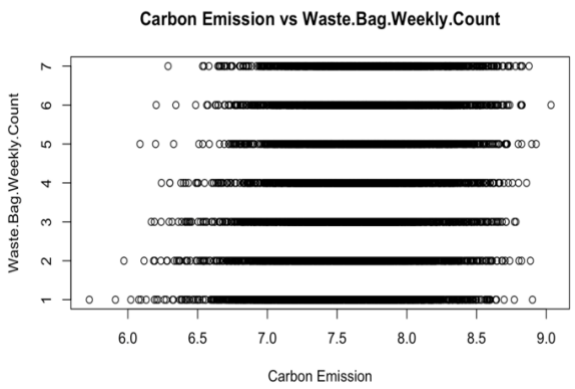

Carbon Emission vs Social.Activity
Social.Activity
6.0 6.5 7.0 7.5 8.0 8.5 9.0
never
often
sometimes
CarbonEmission

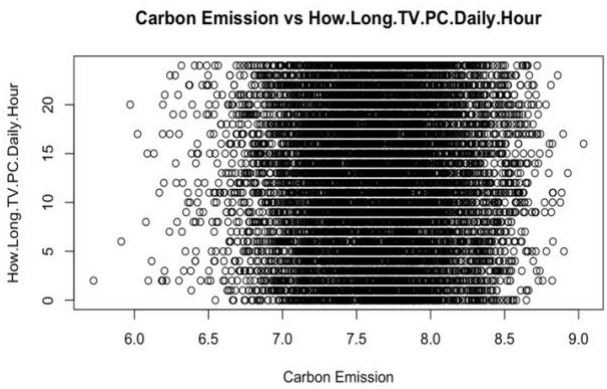

Carbon Emission vs Monthly.Grocery.Bill
Monthly.Grocery.Bill
50 100 150 200 250 300
6.0 6.5 7.0 7.5 8.0 8.5 9.0
Carbon Emission

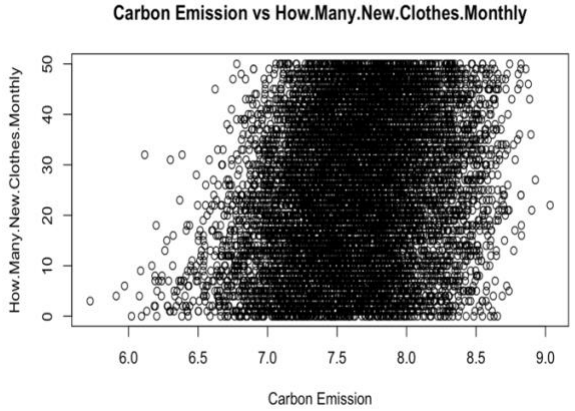

Carbon Emission vs Frequency.of.Traveling.by.Air
Frequency.of.Traveling.by.Air
6.0 6.5 7.0 7.5 8.0 8.5 9.0
frequently
never
rarely
very frequently
CarbonEmission

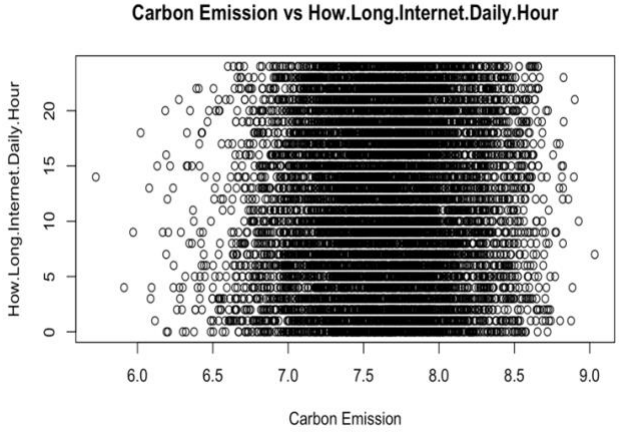

Carbon Emission vs Vehicle.Monthly.Distance.Km
Vehicle.Monthly.Distance.Km
0 2000 6000 10000
6.0 6.5 7.0 7.5 8.0 8.5 9.0
Carbon Emission

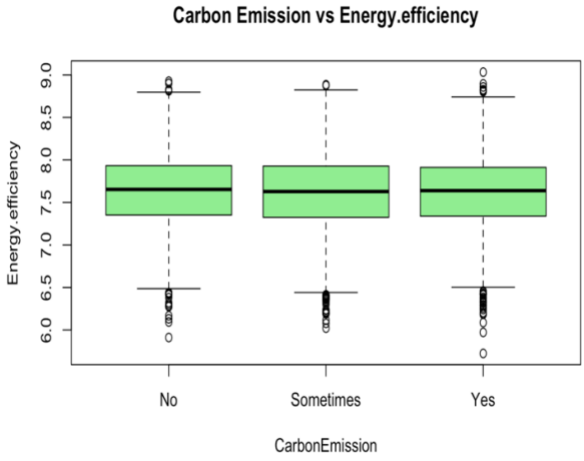

Carbon Emission vs Waste.Bag.Size
Waste.Bag.Size
6.0 6.5 7.0 7.5 8.0 8.5 9.0
extra large
large
medium
small
CarbonEmission

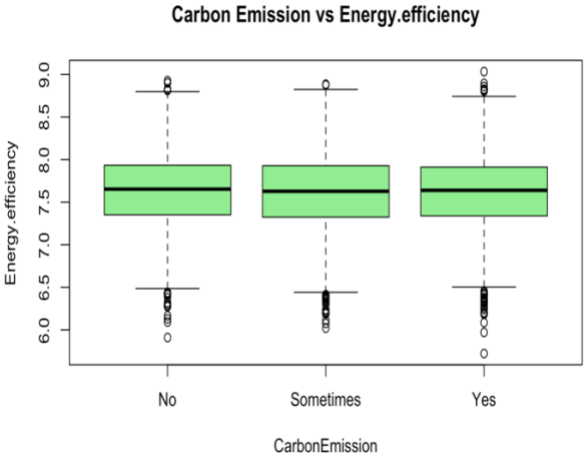

Carbon Emission vs Waste.Bag.Weekly.Count
Waste.Bag.Weekly.Count
Carbon Emission

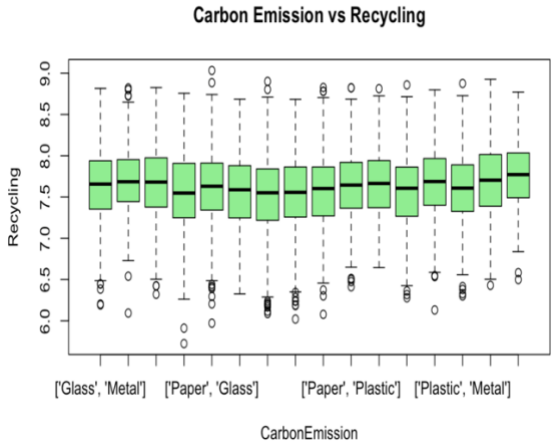

Carbon Emission vs How.Long.TV.PC.Daily.Hour
How.Long.TV.PC.Daily.Hour
Carbon Emission

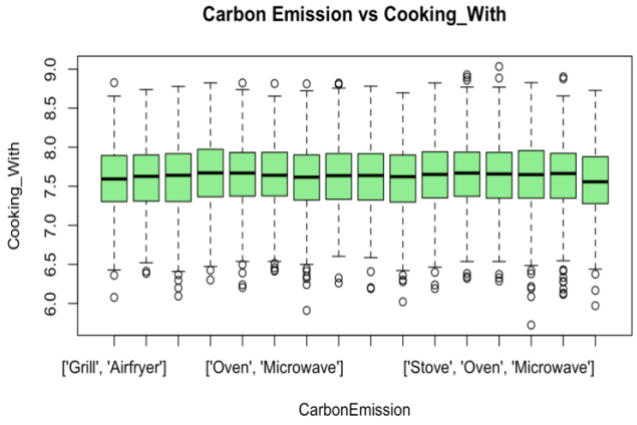

Carbon Emission vs How.Many.New.Clothes.Monthly
How.Many.New.Clothes.Monthly
Carbon Emission


Carbon Emission vs How.Long.Internet.Daily.Hour
How.Long.Internet.Daily.Hour
Carbon Emission


Carbon Emission vs Energy.efficiency
Energy.efficiency
No
Sometimes
Yes
CarbonEmission


Carbon Emission vs Energy.efficiency
Energy.efficiency
No
Sometimes
Yes
CarbonEmission


Carbon Emission vs Recycling
Recycling
6.0
6.5
7.0
7.5
8.0
8.5
9.0
['Glass', 'Metal']
['Paper', 'Glass']
['Paper', 'Plastic']
['Plastic', 'Metal']
CarbonEmission


Carbon Emission vs Cooking_With
Cooking_With
6.0
6.5
7.0
7.5
8.0
8.5
9.0
['Grill', 'Airfryer']
['Oven', 'Microwave']
['Stove', 'Oven', 'Microwave']
CarbonEmission

DATA ANALYSIS AND RESULT

For this dataset we will be using the following models.

- **KNN**
- **Linear Regression and with Backward selection**
- **Lasso**
- **Ridge**
- **Elastic Net Linear**
- **Random Forest**
- **Gradient Boosted Model**
- **Support Vector Machine Linear**
- **Support Vector Machine Radial**
- **Neural Network**

KNN (K-NEAREST NEIGHBORS) MODEL

For the KNN model I first separated my data into to sets i.e. training set and test set and then trained KNN model on the training set and got the following outcome.

```
k-Nearest Neighbors

8001 samples
  26 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6401, 6401, 6401, 6401, 6400
Resampling results across tuning parameters:

  k  RMSE       Rsquared   MAE
  5  0.3918564  0.2337468  0.3118305
  7  0.3852271  0.2464959  0.3062944
  9  0.3813235  0.2552228  0.3027669

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 9.
```

The KNN model output displays the training process used 5-fold cross validation and the smallest subset used for training had 6400 samples and we can see the results of multiple values of k and the RMSE, Rsquared, MAE outcomes for each value of k. The best value of K was determined by the RMSE with the lowest error achieved at k.

LINEAR REGRESSION MODEL

From the output of the linear regression model, we can see that we got RMSE as 2.4018 and Rsquared as 0.9448 and MAE of 1.6809. The Rsquared value suggest that most predictors variable are useful in explaining this variability. The residual error at 2.395 shows that the model prediction derivate from actual emission by an average of 2.395.

```
Linear Regression

8001 samples
  26 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6401, 6400, 6401, 6401, 6401
Resampling results:

  RMSE      Rsquared   MAE
  2.401822  0.9447826  1.680948

Tuning parameter 'intercept' was held constant at a value of TRUE
 Call:
 lm(formula = .outcome ~ ., data = dat)

 Residuals:
      Min       1Q   Median       3Q      Max
 -13.7064  -1.1429   0.0691   1.1851  10.9277

 Coefficients: (2 not defined because of singularities)
                                          Estimate Std. Error t value Pr(>|t|)
 (Intercept)                              4.061e+01  2.298e-01 176.700  < 2e-16 ***
 Body.Typeobese                           4.295e+00  7.593e-02  56.561  < 2e-16 ***
 Body.Typeoverweight                      2.135e+00  7.607e-02  28.071  < 2e-16 ***
 Body.Typeunderweight                    -1.236e+00  7.564e-02 -16.336  < 2e-16 ***
 Sexmale                                  3.486e+00  5.368e-02  64.943  < 2e-16 ***
 Dietpescatarian                         -9.725e-01  7.559e-02 -12.865  < 2e-16 ***
 Dietvegan                               -1.733e+00  7.612e-02 -22.761  < 2e-16 ***
 Dietvegetarian                          -1.519e+00  7.616e-02 -19.939  < 2e-16 ***
 `How.Often.Showerless frequently`       -1.613e-01  7.589e-02  -2.125  0.03361 *
 `How.Often.Showermore frequently`        3.694e-01  7.614e-02   4.852 1.25e-06 ***
 `How.Often.Showertwice a day`            2.304e-01  7.555e-02   3.050  0.00229 **
 Heating.Energy.Sourceelectricity        -4.823e+00  7.533e-02 -64.022  < 2e-16 ***
 `Heating.Energy.Sourcenatural gas`      -2.196e+00  7.608e-02 -28.859  < 2e-16 ***
 Heating.Energy.Sourcewood               -2.168e+00  7.590e-02 -28.565  < 2e-16 ***
 Transportpublic                         -4.203e+00  1.338e-01 -31.412  < 2e-16 ***
 `Transportwalk/bicycle`                 -3.235e+00  1.415e-01 -22.855  < 2e-16 ***
 Vehicle.Typeelectric                    -1.221e+01  1.494e-01 -81.723  < 2e-16 ***
 Vehicle.Typehybrid                      -3.903e+00  1.514e-01 -25.774  < 2e-16 ***
 Vehicle.Typelpg                          1.092e+00  1.481e-01   7.372 1.85e-13 ***
 `Vehicle.TypeNo vehicle`                        NA         NA      NA       NA
 Vehicle.Typepetrol                       5.236e+00  1.510e-01  34.677  < 2e-16 ***
 Social.Activityoften                     1.829e+00  6.535e-02  27.981  < 2e-16 ***
 Social.Activitysometimes                 7.802e-01  6.590e-02  11.838  < 2e-16 ***
 Monthly.Grocery.Bill                     9.861e-03  3.725e-04  26.475  < 2e-16 ***
 Frequency.of.Traveling.by.Airnever      -7.482e+00  7.594e-02 -98.526  < 2e-16 ***
 Frequency.of.Traveling.by.Airrarely     -5.158e+00  7.578e-02 -68.071  < 2e-16 ***
 `Frequency.of.Traveling.by.Airvery frequently`  6.489e+00  7.557e-02  85.868  < 2e-16 ***
 Vehicle.Monthly.Distance.Km              1.841e-03  1.596e-05 115.404  < 2e-16 ***
 Waste.Bag.Sizelarge                     -1.398e+00  7.602e-02 -18.392  < 2e-16 ***
 Waste.Bag.Sizemedium                    -2.875e+00  7.595e-02 -37.851  < 2e-16 ***
 Waste.Bag.Sizesmall                     -4.296e+00  7.586e-02 -56.636  < 2e-16 ***
 Waste.Bag.Weekly.Count                   9.025e-01  1.348e-02  66.960  < 2e-16 ***
 How.Long.TV.PC.Daily.Hour                2.758e-02  3.781e-03   7.295 3.27e-13 ***
 How.Many.New.Clothes.Monthly             1.520e-01  1.818e-03  83.617  < 2e-16 ***
 How.Long.Internet.Daily.Hour             7.945e-02  3.682e-03  21.575  < 2e-16 ***
 Energy.efficiencySometimes              -3.705e-01  6.571e-02  -5.637 1.78e-08 ***
 Energy.efficiencyYes                    -6.688e-01  6.652e-02 -10.055  < 2e-16 ***
 Metal                                   -1.446e+00  5.380e-02 -26.874  < 2e-16 ***
 Paper                                   -1.620e+00  5.370e-02 -30.164  < 2e-16 ***
 Plastic                                 -6.891e-01  5.370e-02 -12.832  < 2e-16 ***
 Glass                                   -1.041e+00  5.369e-02 -19.383  < 2e-16 ***
 Stove                                    2.814e-01  5.368e-02   5.241 1.64e-07 ***
 Oven                                     4.164e-01  5.372e-02   7.750 1.03e-14 ***
 Microwave                                1.153e-01  5.364e-02   2.149  0.03164 *
 Grill                                    3.887e-01  5.366e-02   7.244 4.77e-13 ***
 Airfryer                                        NA         NA      NA       NA
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 2.395 on 7957 degrees of freedom
 Multiple R-squared:  0.9453,    Adjusted R-squared:  0.945
 F-statistic:  3200 on 43 and 7957 DF,  p-value: < 2.2e-16
```

Body type was a significant predictor where people with “obese” category increased the emission by 4.3 units and individual who were in “underweight” category reduced the emission by 1.2 units. **Vehicle type**, people who used electric vehicle by decreased the emission by over 12 units and other type of vehicles like hybrid and petrol increased the emission. **Vehicle Monthly Distance Km** it shows that for every km traveled emission increased proportionally. From the overall it shows that predictor variables collectively have an impact on predicting carbon emission.

STEPWISE LINEAR REGRESSION MODEL

This linear regression model was built using backward selection. It is a technique that sequentially removes predictors to find the combination that minimizes prediction error. From the model’s performance we got RMSE as 7.6546, Rsquared as 0.4389 and MAE as 6.1349 indicating the predictors has variability in carbon emission.

```
Linear Regression with Backwards Selection

8001 samples
  26 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6401, 6402, 6401, 6400, 6400
Resampling results across tuning parameters:

  nvmax  RMSE      Rsquared   MAE
  2      8.431435  0.3193071  6.764478
  3      7.960447  0.3929615  6.451064
  4      7.654569  0.4388670  6.134912

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was nvmax = 4.
```

The final predictors that got selected are **Vehicle type – Electric** with low carbon emission compared to other types. **Frequency of Traveling by Air – Very frequently** contributed to higher emission. **Vehicle monthly Distance Km** more kilometers traveled by vehicle are positively corelated in increasing the carbon emissions.

```
1 subsets of each size up to 5
Selection Algorithm: backward
         Body.Typeobese Body.Typeoverweight Body.Typeunderweight Sexmale Dietpescatarian Dietvegan Dietvegetarian How.Often.Showerless frequently
1  ( 1 ) " "            " "                 " "                 " "     " "              " "        " "             " "
2  ( 1 ) " "            " "                 " "                 " "     " "              " "        " "             " "
3  ( 1 ) " "            " "                 " "                 " "     " "              " "        " "             " "
4  ( 1 ) " "            " "                 " "                 " "     " "              " "        " "             " "
5  ( 1 ) " "            " "                 " "                 " "     " "              " "        " "             " "
         How.Often.Showermore frequently How.Often.Showertwice a day Heating.Energy.Sourceelectricity Heating.Energy.Sourcenatural gas
1  ( 1 ) " "                             " "                         " "                              " "
2  ( 1 ) " "                             " "                         " "                              " "
3  ( 1 ) " "                             " "                         " "                              " "
4  ( 1 ) " "                             " "                         " "                              " "
5  ( 1 ) " "                             " "                         " "                              " "
         Heating.Energy.Sourcewood Transportpublic Transportwalk/bicycle Vehicle.Typeelectric Vehicle.Typehybrid Vehicle.Typelpg
1  ( 1 ) " "                       " "             " "                  " "                  " "                " "
2  ( 1 ) " "                       " "             " "                  " "                  " "                " "
3  ( 1 ) " "                       " "             " "                  "*"                  " "                " "
4  ( 1 ) " "                       " "             " "                  "*"                  " "                " "
5  ( 1 ) " "                       " "             " "                  "*"                  " "                " "
         Vehicle.TypeNo vehicle Vehicle.Typepetrol Social.Activityoften Social.Activitysometimes Monthly.Grocery.Bill
1  ( 1 ) " "                    " "                " "                  " "                     " "
2  ( 1 ) " "                    " "                " "                  " "                     " "
3  ( 1 ) " "                    " "                " "                  " "                     " "
4  ( 1 ) " "                    " "                " "                  " "                     " "
5  ( 1 ) " "                    " "                " "                  " "                     " "
         Frequency.of.Traveling.by.Airnever Frequency.of.Traveling.by.Airrarely Frequency.of.Traveling.by.Airvery frequently
1  ( 1 ) " "                                " "                                 " "
2  ( 1 ) " "                                " "                                 "*"
3  ( 1 ) " "                                " "                                 "*"
4  ( 1 ) " "                                " "                                 "*"
5  ( 1 ) "*"                                " "                                 "*"
         Vehicle.Monthly.Distance.Km Waste.Bag.Sizelarge Waste.Bag.Sizemedium Waste.Bag.Sizesmall Waste.Bag.Weekly.Count
1  ( 1 ) "*"                         " "                 " "                  " "                 " "
2  ( 1 ) "*"                         " "                 " "                  " "                 " "
3  ( 1 ) "*"                         " "                 " "                  " "                 " "
4  ( 1 ) "*"                         " "                 " "                  " "                 " "
5  ( 1 ) "*"                         " "                 " "                  " "                 " "
         How.Long.TV.PC.Daily.Hour How.Many.New.Clothes.Monthly How.Long.Internet.Daily.Hour Energy.efficiencySometimes Energy.efficiencyYes
1  ( 1 ) " "                       " "                         " "                          " "                       " "
2  ( 1 ) " "                       " "                         " "                          " "                       " "
3  ( 1 ) " "                       " "                         " "                          " "                       " "
4  ( 1 ) " "                       "*"                         " "                          " "                       " "
5  ( 1 ) " "                       "*"                         " "                          " "                       " "
         Metal Paper Plastic Glass Stove Oven Microwave Grill Airfryer
1  ( 1 ) " "   " "   " "     " "   " "  " "  " "       " "   " "
2  ( 1 ) " "   " "   " "     " "   " "  " "  " "       " "   " "
3  ( 1 ) " "   " "   " "     " "   " "  " "  " "       " "   " "
4  ( 1 ) " "   " "   " "     " "   " "  " "  " "       " "   " "
5  ( 1 ) " "   " "   " "     " "   " "  " "  " "       " "   " "
```

RANDOM FOREST MODEL

In the random forest model, the model's performance shows mean of squared residuals at 0.0164. the high variance indicates that the random forest model can account over 91% of the variability in carbon emission.

```
Call:
 randomForest(formula = CarbonEmission ~ ., data = carbonTrainData)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 8

          Mean of squared residuals: 0.01640799
                    % Var explained: 91.45
```

The cross-validation results show us that the best preforming model used 23 predictors at each split mtry=23 and got RMSE 0.1386710, Rsquared 0.9037716, MAE 0.10667

```
Random Forest

8001 samples
  26 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6400, 6400, 6402, 6401, 6401
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE
   2    0.2681786  0.8113335  0.2093666
  23    0.1386710  0.9037716  0.1066725
  45    0.1428352  0.8953120  0.1101724

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 23.
rf variable importance

  only 20 most important variables shown (out of 45)
```

In the variable important analysis, it showed that **vehicle monthly distance km** was the most influential predictor. Frequency of travelling by air (Very Frequently) was strongly co related with higher emission and other were even related with it.

	Overall <dbl>
Vehicle.Monthly.Distance.Km	100.000000
Frequency.of.Traveling.by.Airvery frequently	75.821118
Vehicle.Typeelectric	35.747772
Frequency.of.Traveling.by.Airnever	33.554032
How.Many.New.Clothes.Monthly	32.791992
Vehicle.TypeFuelEfficient	20.919024
Frequency.of.Traveling.by.Airrarely	17.759897
Waste.Bag.Weekly.Count	17.109349
Heating.Energy.Sourceelectricity	10.533998
Sexmale	8.838146

Random forest model effectively identifies and ranks the most influential carbon emission. The Cross-validation results indicate that the model's predictive power for understanding the behavioral determinants of emission.

GRADIENT BOOSTED MODEL

In the gradient boosted model the optimal model was found to have 150 trees and an interaction depth of 3 with a shrinkage rate 0.1 and n.minobsinnode 10 and the RMSE of 0.1100, Rsquared of 0.9440 and MAE of 0.0816 demonstrating that it can accurately predict carbon emissions based on the given predictors.

```
Stochastic Gradient Boosting

8001 samples
  26 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6400, 6402, 6401, 6400, 6401
Resampling results across tuning parameters:

  interaction.depth  n.trees  RMSE       Rsquared   MAE
  1                   50      0.2821208  0.6764349  0.21893567
  1                  100      0.2268785  0.7947714  0.17367395
  1                  150      0.1956806  0.8372088  0.14803391
  2                   50      0.2169693  0.8219946  0.16569193
  2                  100      0.1621487  0.8899824  0.12168082
  2                  150      0.1357390  0.9167269  0.10113605
  3                   50      0.1879962  0.8612388  0.14324246
  3                  100      0.1355744  0.9205363  0.10204167
  3                  150      0.1100162  0.9440393  0.08159571

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at a value of 10
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.
```

SUPPORT VECTOR LINEAR MODEL

The SVM with linear kernel output shows good predictive performance the model approx. has 92.35% of the variance in the carbon emission data. And has the lowest RMSE and MAE which indicates accurate predications.

```
Support Vector Machines with Linear Kernel

8001 samples
  26 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6400, 6402, 6401, 6400, 6401
Resampling results:

  RMSE       Rsquared  MAE
  0.1215561  0.923485  0.08597153

Tuning parameter 'C' was held constant at a value of 1
```

SUPPORT VECTOR RADIAL MODEL

The SVM with Radial basis function kernel demonstrates the optimal model was selected based on the lowest RMSE which is 0.0601 the final parameter values used are C is1.0 and sigma as 0.1201559. This model explains approximately 98.17% of the variance in the carbon emission data and lowest RMSE and MAE.

```
Support Vector Machines with Radial Basis Function Kernel

8001 samples
  26 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6400, 6402, 6401, 6400, 6401
Resampling results across tuning parameters:

  C     RMSE        Rsquared   MAE
  0.25  0.07973153  0.9690666  0.05509160
  0.50  0.06746759  0.9773167  0.04740922
  1.00  0.06008418  0.9816922  0.04317555

Tuning parameter 'sigma' was held constant at a value of 0.01201559
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were sigma = 0.01201559 and C = 1.
```

COMPARING MODEL'S PERFORMANCE

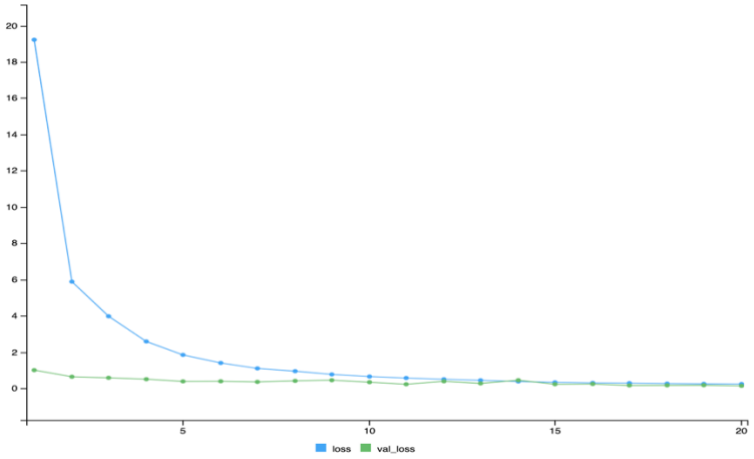
From all the models (KNN, LIN, Stepwise, Lasso, Ridge, Enet, RF, GBM, SVML, SVMR) we have got out all the models RMSE, Rsquared and MAE.

```
Call:
summary.resamples(object = compare)

Models: KNN, LIN, stepWise, Lasso, Ridge, Enet, RF, GBM, SVML, SVMR
Number of resamples: 5

MAE
               Min.    1st Qu.     Median       Mean    3rd Qu.       Max. NA's
KNN      0.29557166 0.30331883 0.30555426 0.30388053 0.30648972 0.30846817    0
LIN      0.08387579 0.08609215 0.08621809 0.08673289 0.08664359 0.09083484    0
stepWise 0.26929668 0.27271591 0.27310626 0.27362039 0.27648573 0.27649740    0
Lasso    0.08610631 0.08642567 0.08703662 0.08693581 0.08723702 0.08787345    0
Ridge    0.08746251 0.08798745 0.08805830 0.08854089 0.08888671 0.09030947    0
Enet     0.08603446 0.08605687 0.08696593 0.08676127 0.08723457 0.08751452    0
RF       0.10442811 0.10645550 0.10691508 0.10667247 0.10693527 0.10862838    0
GBM      0.08012308 0.08043568 0.08118687 0.08159571 0.08273428 0.08349863    0
SVML     0.08459812 0.08542030 0.08623506 0.08597153 0.08665323 0.08695093    0
SVMR     0.04269283 0.04279000 0.04285568 0.04317555 0.04351738 0.04402187    0

RMSE
               Min.    1st Qu.     Median       Mean    3rd Qu.       Max. NA's
KNN      0.36923040 0.37985702 0.3851089 0.38198895 0.38571449 0.39003388    0
LIN      0.11804858 0.11867965 0.1192947 0.12107146 0.12055212 0.12878231    0
stepWise 0.34293536 0.34454591 0.3449272 0.34679250 0.35023234 0.35132167    0
Lasso    0.11780458 0.11944983 0.1220385 0.12144810 0.12315524 0.12479239    0
Ridge    0.12004312 0.12079011 0.1248700 0.12347245 0.12558131 0.12607774    0
Enet     0.11781611 0.11919223 0.1215651 0.12120694 0.12260288 0.12485842    0
RF       0.13605544 0.13809940 0.1386317 0.13867100 0.13906601 0.14150247    0
GBM      0.10724642 0.10767185 0.1094101 0.11001615 0.11006181 0.11569063    0
SVML     0.11734991 0.11959173 0.1221653 0.12155609 0.12328870 0.12538475    0
SVMR     0.05730271 0.05964589 0.0597342 0.06008418 0.06184576 0.06189236    0

Rsquared
              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
KNN      0.2296715 0.2319757 0.2440333 0.2505040 0.2636345 0.2832051    0
LIN      0.9198126 0.9223686 0.9251159 0.9237080 0.9255435 0.9256996    0
stepWise 0.3451496 0.3526628 0.3675152 0.3731130 0.3959060 0.4043313    0
Lasso    0.9199537 0.9215727 0.9245300 0.9233747 0.9253924 0.9254250    0
Ridge    0.9189336 0.9209425 0.9237481 0.9224667 0.9242417 0.9244679    0
Enet     0.9205147 0.9215462 0.9248279 0.9236046 0.9253255 0.9258086    0
RF       0.9018731 0.9020678 0.9034705 0.9037716 0.9055442 0.9059024    0
GBM      0.9389176 0.9433262 0.9437315 0.9440393 0.9464198 0.9478012    0
SVML     0.9200256 0.9212304 0.9246664 0.9234850 0.9256671 0.9258354    0
SVMR     0.9809750 0.9814003 0.9815973 0.9816922 0.9818137 0.9826749    0
```

From all the models we can see that SVMR has got the lowest MAE (0.0432), Lowest RMSE which is (0.0601) and highest Rsquared (0.9817). Overall SVMR stands out over all the metrics, whereas GBM and LIN also had high predictivity, but SVMR outperformed.

NEURAL NETWORK

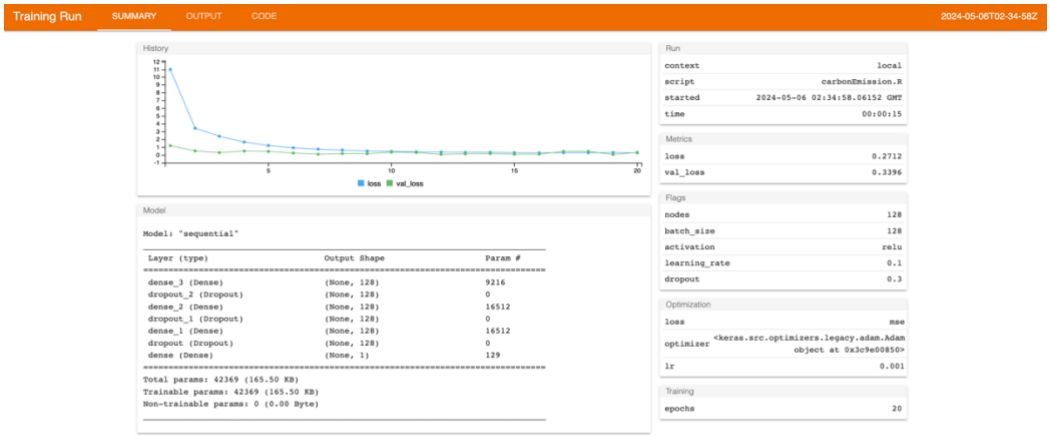
For the neural network I used 3 input layers and activation as RELU as I am making an regression model I didn't add any activation for the output layer with which I got the following graph output.



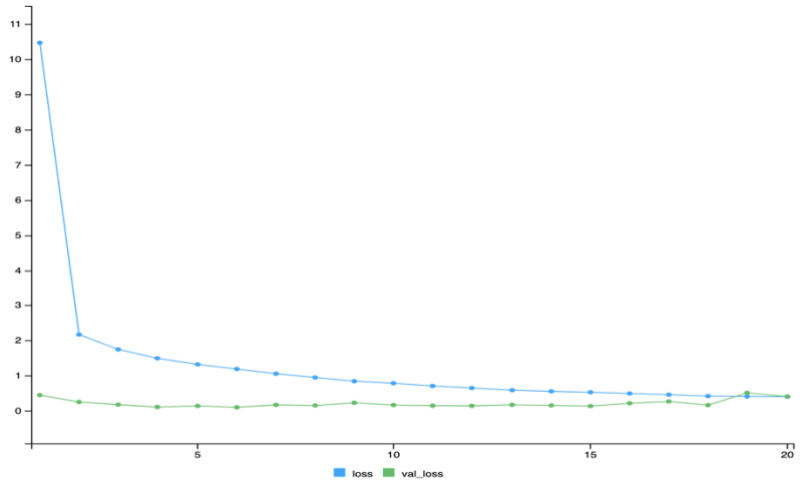
After calculating the RMSE, MAE and the Rsquared value I got the following output.

```
[1] 0.3150746
[1] 0.3283859
[1] 0.6970954
```

After fine tuning, I selected the best hyperparameters and combined my training set and validation set and then passed the test set and got the following results.



This is the output of my best hyperparameter.



After using the hyperparameter we can see that the training loss steadily decreased, and the validation loss kept fluctuating and stabilized at the end. After calculating the RMSE the model's predictions deviated by approximately 0.3151, MAE showed that the predictions are off by 0.3284 and the Rsquared came out 69.71% of the variability in the target variable. However, it still lags, some other models like SVM Radial and the GRB in terms of RMSE and Rsquared.

```
[1] 0.3150746
[1] 0.3283859
[1] 0.6970954
```

CONCLUSION

In this project we explored various machine learning models to predict carbon emission based on the lifestyle factors. Lifestyle factors like transportation, diet, and energy source hold significant predictive power in emission of carbon this helps in finding that daily habits and choices can contribute to an individual environmental footprint.

We saw that various models show varying capabilities in capturing the complex relationship between lifestyle and carbon emission. From all the models we saw that Support Vector Machine and Gradient Boosting Machine theoretically prove to be powerful due to their ability to model nonlinear parameters, which are essential to the environment.

We even got to know that complex model such as neural network can theoretically capture suitable pattern, their interpretability is limited while compared to simple linear models.

Overall, we can say that machine learning models can be used to predict and understand the impact of lifestyle choices on carbon emission.

REFERENCES

https://www.kaggle.com/datasets/dumanmesut/individual-carbon-footprint-calculation/data